Bases de datos avanzadas

Fèlix Saltor
Catedrático de Lenguajes y Sistemas Informáticos
Universitat Politècnica de Catalunya

Acceso integrado a bases de datos Heterogeneas*

1. Sistemas de bases de datos federadas

1.1 Motivación

Supongamos que existen varias bases de datos, cada una de las cuales ha surgido independientemente de las otras, por lo que ha sido diseñada y es gestionada con total autonomía respecto a las demás. Si en estas condiciones aparece la necesidad, para un usuario o grupo de usuarios, de acceder al conjunto de los datos mantenidos en todas estas base de datos como un todo integrado, como una sola bases de datos, hay dos caminos para conseguirlo:

a) Integración. Diseñar y construir una nueva base de datos, en la que se integren los contenidos de las preexistentes, con la consiguiente conversión de las antiguas a la nueva. La conversión a la nueva base de datos de los programas y métodos de trabajar de los usuarios preexistentes será más o menos costosa según los casos. Es posible que las bases de datos preexistentes residieran en distintas localidades o en edificios diferentes; puede ser adecuado que la nueva base de datos sea distribuida.

b) Federación. Mantener las bases de datos preexistentes tal y como estaban, y superponer un nuevo sistema sobre sus Sistemas de Gestión de Bases de Datos (SGBD). Este sistema presenta el conjunto de las bases de datos, como si se tratase de una sola base de datos, a los usuarios que lo requieren: el usuario formula una sola pregunta al sistema, y este devuelve una sola respuesta (en el proceso de la consulta y en la confección de la respuesta interoperan las bases de datos preexistentes, aunque ésto es transparente al usuario). Por otra parte, los programas y los usuarios preexistentes (de una sola de las bases de datos) no se ven afectados.

Cada vez que surja la necesidad del acceso integrado, habrá que estudiar la factibilidad y la conveniencia de la integración y de la federación, y tomar la decisión adecuada.

1.2 Características

La federación da lugar a un Sistema de Bases de Datos Federadas (SBDF); otros autores hablan de bases de datos interoperables o de multibases. No hay pues integración de las bases de datos preexistentes, que denominaremos bases de datos componentes; sí hay integración del acceso por parte de los usuarios, en contraposición a acceder separadamente a cada una de la bases de datos e integrar manualmente las respectivas respuestas; se dice que hay acceso integrado.

Tenemos pues dos niveles, como mínimo: el nivel de las bases de datos (B de D) componentes, y el nivel federal del SBDF.

Un SBDF se caracteriza por la autonomía y la heterogeneidad de sus bases de datos componentes. Puesto que éstas residen, en general, en instalaciones distintas, el SBDF será distribuido; sin embargo, su distribución es una consecuencia, no una característica, de su definición. Por otra parte, las bases de datos componentes pueden ser distribuidas.

Una base de datos puede ser componente de varios SBDF, las federaciones pueden formarse y desaparecer, y en ellas pueden entrar y salir bases de datos componentes. En general, un SBDF no tiene necesariamente un Esquema de Base de Datos conceptual único, común a toda la federación, sino que puede tener diversos Esquemas en el nivel federal.

1.3 Autonomía

Las bases de datos componentes de un SBDF tienen autonomía:

- de diseño, puesto que han sido concebidas independientemente, y ello da lugar a la heterogeneidad;
- de comunicación, es decir para decidir con quién, qué, y cómo se va a compartir datos (interoperar); y
- de ejecución, sobre cuándo y de qué manera procesar las consultas, de un modo compatible con sus usuarios y aplicaciones preexistentes.

1.4 Heterogeneidad

Como consecuencia de la autonomía, puede existir heterogeneidad:

- de sistemas, de CPU, de Sistema Operativo, de modelo de datos, de Sistema de Gestión de B de D
- de datos, de concepción del mundo real, de modelización de esta concepción del mundo real en términos del modelo de datos y del SGBD

1.5 Arquitectura

Entre las arquitecturas de referencia propuestas para un SBDF, cabe destacar la de Sheth & Larson:

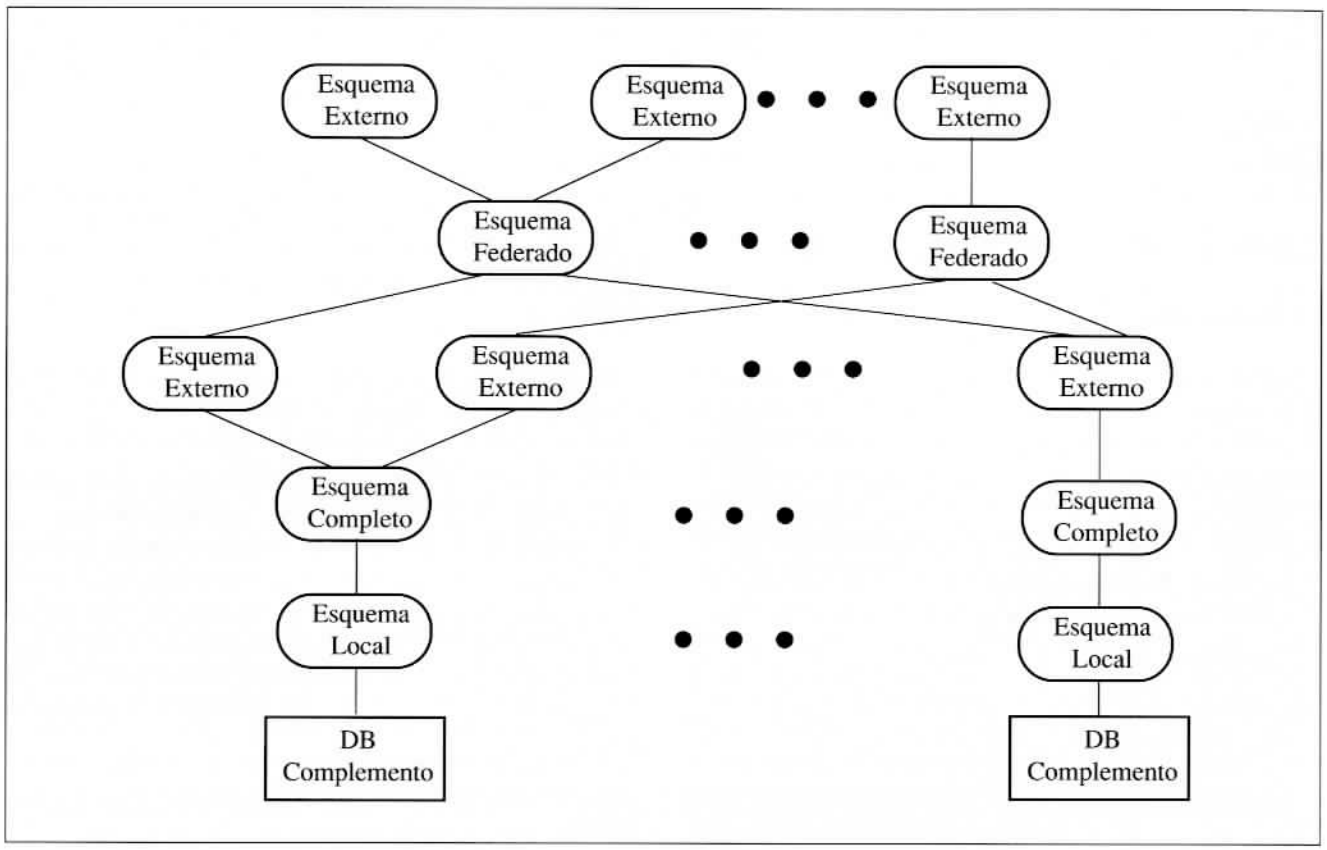

Figura. Arquitectura de Esquemas a 5 niveles de un SBDF

2. Problemática

Para conseguir el acceso integrado de un SBDF hace falta solucionar, entre otros, los siguientes problemas, en el caso más general.

2.1 Negociación de derechos de acceso y seguridad

El Administrador de una B de D componente negocia con un usuario federal (usuario final o administrador federal):

- A qué datos le permite acceder
- En qué modalidades: consulta o actualización
- A cambio de qué (es Administrador de otra B de D componente), o en virtud de qué (legislación)

2.2 Heterogeneidad sintáctica

Hablaremos de heterogeneidad sintáctica en los siguientes casos:

- Heterogeneidad de modelo de datos: bijerárquico, invertido, CODASYL, relacional, uno de los orientados a objetos, etc. Para solucionarlo, se utiliza un modelo canónico común a toda la federación.
- Heterogeneidad de implementación del modelo (SGBD). En el caso relacional: Ingres, Oracle, DB2, Informix, Rdb, Mimer, Unify, Sybase, Supra, Transtools, etc. Hay que establecer correspondencias entre esquemas del SGBD y esquemas del modelo canónico.
- Heterogeneidad de lenguajes. Hay estándares (NDL, SQL), aunque cada implementación difiera del estándar. Hacen falta traducciones del y al lenguaje canónico (o lenguajes).

2.3 Enriquecimiento semántico

Cada Esquema de una B de D componente (almacenado en su catálogo) tiene como nivel semántico el de su propio modelo. El modelo canónico ha de tener un nivel semántico superior o igual al de cada B de D componente.

Hace falta elevar el nivel semántico de los Esquemas de las B de D componentes hasta el nivel semántico del modelo canónico, antes de proceder a integrarlos. Un esquema enriquecido es pues una base de conocimientos. Se está investigando sobre técnicas semiautomáticas de enriquecimiento semántico: por ejemplo, un modelo de relacional a orientado a objetos.

2.4 Integración de esquemas

Para integrar los esquemas enriquecidos, es necesario poder identificar clases y atributos equivalentes, y para ello

superar sus diferencias semánticas: polisemia, homonimia, de dominio, de escala y de unidad de medida.

También hay que proceder a la identificación de individuos, es decir, cuando un mismo individuo aparece en más de una de las bases de datos componentes, identificarlo como un solo individuo y no como varios.

Finalmente, hay que estudiar las interdependencias entre las B de D componentes.

2.5 Descomposición del acceso y composición del resultado

Intervienen tres fases:

a) Descomponer la consulta en subconsultas a las B de D componentes afectadas y traducir cada subconsulta del lenguaje canónico a un lenguaje del SGBD correspondiente.
b) Ejecutar las subconsultas "autónomamente" (ver 2.6).
c) Transformar cada subresultado al modelo canónico, y componerlos para producir el resultado a nivel federal. Ello implica la identificación de individuos.

Si el acceso actualiza, y no sólo consulta, se presenta el problema de las actualizaciones y la atomicidad (ver 2.6).

2.6 Gestión de transacciones

En un SGBD convencional, una transacción tiene las características de: Atomicidad, Consistencia, Aislamiento y Perdurabilidad.

En un Sistema de B de D Federadas, la autonomía es incompatible con la atomicidad; la heterogeneidad de mecanismos de control de concurrencia de los SGBDs hace difícil el aislamiento a escala federal; y los SGBDs no necesariamente tienen primitivas adecuadas para unos protocolos de coordinación a escala federal

En consecuencia, se están investigando otras clases de "transacciones", que también pueden ser útiles en otros contextos.

3. Conclusiones

El acceso integrado a Bases de Datos heterogéneas es un problema cada vez más real.

Establecer sistemas de B de D Federadas requiere, en el caso más general, solucionar una serie de problemas. En casos particulares, alguno de los problemas puede no darse. Por ejemplo, si todas las bases de datos componentes son relacionales y utilizan versiones compatibles de SQL, no hay heterogeneidad sintáctica. Si todos los accesos consultan y no actualizan, la gestión de transaciones queda simplificada. De todos modos, la heterogeneidad semántica se presentará siempre.

Se está investigando sobre estos problemas a escala mundial: en distintas universidades y en consorcios entre instituciones (como el consorcio IHIS, con la participación de las Universidades Politécnicas de Madrid y Cataluña); mediante programas específicos en EEUU y en Europa; en congresos especializados (como el de Kyoto 1991), y con la aparición de números especiales de revistas.

** Una versión de este artículo fue presentada en el Encuentro sobre Bases de Datos en la Administración Pública, organizado por el CREI en Madrid en 1990.*